



# FÖREDRAG

vid

De nordiska  
datalogvistik  
dagarna 1983

Uppsala den 3-4 oktober

Utgivare  
Anna Sågvall Hein

Centrum för datorlingvistik,  
Uppsala universitet

Fred Karlsson  
Department of General Linguistics  
University of Helsinki  
Hallituskatu 11-13  
SF-00100 HELSINKI 10  
FINLAND

## TAGGING AND PARSING FINNISH

### 1. Introductory remarks

Current parsing theory has an obvious English bias. Many of the problems discussed originate in typical features of English such as the scarcity of surface morphosyntactic markers, the scarcity of word forms, and the occurrence of certain types of ambiguous syntactic structures. E.g. Winograd (1983, 544-5) argues that morphological phenomena do not lend themselves well to the methodology of generative grammar because of their high degree of irregularity and idiosyncrasy. Furthermore, he opines that any analysis seeking to find morphological regularities must examine words in terms of their history. He even goes so far as to argue that, in contrast to what holds for syntax, native speakers do not utilize grammatical knowledge in the production and understanding of morphological structures. He concedes, though, that there are a few highly productive morphological phenomena that cannot be handled by lexical look-up. But the bulk of morphology is anyway to be deposited in the lexicon just by listing individual forms.

This view does, perhaps, descriptive justice to large portions of English morphology even though the demarcation line between syntax and morphology seems unnecessarily strict even here. But a general model of morphological competence and processing must surely provide stronger means for dealing with e.g. the plethora of word forms found in more synthetic languages.

This paper sketches some basic problems to be dealt with in tagging and parsing synthetic, morphologically rich languages such as Finnish. We see tagging and parsing as closely interdependent (cf. Brodda 1982). A good tagging program is equivalent to a parser in important respects. It is also a methodological prerequisite to doing large-scale parsing, which cannot succeed without the possibility of checking hypotheses on easily accessible, large, grammatically coded corpora.

It is self-evident that the wealth of overt ending morphs is beneficial e.g. in determining constituent structure, dependencies, and syntactic functions. A sufficiently general morphological analyzer conclusively solves most local syntactic problems and provides valuable clues to long-distance dependencies as well. Syntactic parsing strategies of the types discussed e.g. by Bever (1970) and Kimball (1973) certainly have to be widely employed in parsing synthetic languages. But from a general point of view, we see as particularly important the make-up of the lexicon that emerges when one tries to model intricate systems of inflexion and derivation.

Specifically, this paper deals with derivational morphology and morphological tagging as initial phases of a project with the aim of constructing a morphosyntactic parser for unrestricted Finnish text. The final system should i.a. be able to cope with the **whole vocabulary** including exceptions, loan-words, neologisms, potential productive derivatives, etc. We see it as important not to compromise in this respect since initial restrictions to micro-worlds or vocabularies consisting of only a few hundred words tend to severely misrepresent the problems encountered when the whole lexicon is faced. The system should optimally be **bidirectional**, i.e. able both to analyze and to synthesize (produce) word forms and sentences. It should also, as far as possible, have an interpretation as a model of the "real" morphosyntactic processes. No semantics has yet been included. We take one minimal requirement of an adequate semantics to be the ability of the system to treat all morphosyntactic forms.

## 2. A model of derivational morphology

Koskenniemi (1983; also cf. this volume) has designed a general model for word-form recognition and production. The model has been applied to Finnish and yielded a full description of inflexional morphology satisfying the requirements outlined above. The current running lexicon contains the top 3,000 entries of the Finnish frequency dictionary. The model analyzes and produces all the (inflexional and cliticized) 2,000 forms of Finnish nominal paradigms and the 18,000 forms of verb paradigms. The ultimate lexicon will contain some 10,000 entries which suffices for analyzing ordinary running text (excluding, of course, specialized vocabulary).

Koskenniemi's model also analyzes **compounds** provided the (base forms of their) constituent parts are in the lexicon. Compounding is such a central morphological means in synthetic languages that it must be easily tractable also in computational models aspiring general applicability.

The third morphological domain to be covered is **derivation**. The total number of derivational morphemes in Finnish is 150-200 depending upon how the most opaque and unfrequent ones are interpreted. Some 50-60 are highly productive and these are to be discussed here. The maximal productive Finnish derivational system comprises nine morphotactic positions with the following contents (Karlsson 1983).

1 2 3 4 5 6 7 8 9  
(root) V<sub>1</sub> V<sub>2</sub> V<sub>3</sub> PASS N<sub>1</sub> N<sub>2</sub> A<sub>1</sub> A<sub>2</sub> N<sub>3</sub> (infl.) (clit.)

Positions 1,2,3 contain endings deriving verbs from either nouns or verbs. These endings are mainly causative, reflexive, frequentative, or momentaneous. Combinations of up to three of these occur (e.g. lue/t/utt/ele 'read/caus/caus/freq', i.e. 'habitually make somebody to read'). Position 4 has only one member, the so-called passive (better: indefinite). Position 5 contains the majority of (denominal or deverbal) noun endings, position 6 a few noun endings appending to those in position 5 (e.g. asu/nto/la 'hostel'). Position 7 contains the majority of

adjectival endings, position 8 a few adjectival endings ap-  
pending to position 7 (e.g. kiihty/vä/mpi 'more accelerating').  
Position 9, finally, contains only one ending deriving nouns  
from derived adjectives (kiihty/vä/mm/yys 'the property of  
being more accelerating').

The derivational system is further complicated by the  
possibility of repeated recursion from positions 7 and 9, i.e.  
some derived adjectives and nouns permit further derivation of  
verbs etc during a second and even third pass through the  
morphotactic flow chart. This provides for derivatives such as  
oike/ude/llis/ta/minen (approximately:) 'causing to be legal',  
where ude is  $N_3$  (first pass), llis is  $A_1$  (second pass), ta is  
 $V_1$  (third pass) and minen is  $N_1$  (third pass).

As a first step, the derivational system has been modelled  
as a BETA rule-system (cf. Brodda (forthcoming), Brodda &  
Karlsson 1981) that generates all productive derivatives for  
any given input root. The morphophonological dependencies be-  
tween roots and endings, and between endings in sequences, have  
been precisely described and implemented as BETA-rules in order  
to prevent at least morphological overgeneralizations. This  
descriptive task alone is a considerable venture. Semantic  
restrictions, however, have not yet been considered.

We pick an example. Appendix 1 contains the near-maximal  
set of 161 derivatives, including some awkward ones, that the  
BETA rule system generates for the verb hakkaa 'hew'. The first  
six forms (group a) are derived verbs with flow chart positions  
1,2,3 and some combinations of these filled. Then follow (group  
b) 57 nominal derivatives based on the underived root, in-  
cluding combinations of up to three nominal endings. Then  
follow minimal sets (c-h, 16-17 members in each) of nominal  
derivatives of the derived verbs (a). The sets (c-h) are  
conservatively composed. If made maximal, they would contain  
more than 50 members each.

Now recall that each of the six derived verbs has some  
18,000 inflexional and cliticized forms, and all the 155 de-  
rived nominals some 2,000 such forms. This would give an  
astounding sum total of more than 400,000 potential productive  
derived, inflected, and cliticized forms of the single verb

hakkaa.

Such figures clearly show the untenability of any attempt to treat the bulk of synthetic morphological systems, be they agglutinative as Turkish or semi-agglutinative as Finnish, by way of mere lexical listing. The vast majority of the 400,000 forms under scrutiny certainly are both morphologically and semantically regular in the strongest sense of the word, i.e. their morphophonological behaviour and compositional meaning is predictable. These forms must be generated by rules in autonomous grammars as well as in models of morphological performance. Of course, this is not to draw the demarcation line between rules and lexicon where it was drawn by SPE-type phonology. We just stress that the huge majority of forms in synthetic languages must be described as products of rules. Several recent anglocentric theories of the lexicon have gone too far in the opposite direction by including all or most forms in the lexicon (cf. Lieber 1981 for references). Also note that it would make no psycholinguistic sense to claim that the Finnish lexicon is tens, or even hundreds, of thousands of times larger than the English one.

The BETA rule system provides an initial formalization in the process mode of Finnish derivational morphology. The following step will be to incorporate these productive derivational mechanisms in the Finnish implementation of Koskenniemi's two-level model. This makes it possible to reduce the size of the lexicon by eliminating all morphologically and semantically predictable derivatives. It also vastly improves the possibilities of the system to deal with running text. Note e.g. that the system then is able to analyze also occasional compounds with (any number of) derived constituents. It will, as a case in point, have the full power of coping with all the 400,000 forms of the verb hakkaa, both as independent words and as constituents in compounds.

This is the kind of full coverage needed in any reliable practical application such as information retrieval or spelling correction.

### 3. Morphological tagging: FINTAG

Readily available facilities for testing hypotheses and alternatives on large natural corpora are mandatory when one tries to design a parser for running text. For this purpose we have devised a morphological tagging program, FINTAG, that consists of thirteen BETA rule modules geared to apply in sequence to any input text. The output is the original text so analyzed that (a) all word forms have been tagged with a part of speech label, and (b) all inflexional endings and clitics have been segmented. The tagging format is intagging in the sense of Brodda (1982), i.e. the part of speech labels are prefixed to the word forms (separated by a colon). The output format is thus e.g. PR:TÄMÄ=N N:VUODE=N N:ALKU VF:ON A:KYLMA=Ä N:AIKA=A 'the beginning of this year is a cold time', where the tags have standard meanings (VF = finite verb), =N is the genitive sg. ending, and =Ä, =A are partitive sg. endings. The equation mark (=) serves as ending juncture except in illatives (§) and possessives (").

The thirteen BETA rule modules (containing a total of some 7000 lines of substitution rules) have been mutually sequenced both on linguistic and strategic grounds. Some of the rule modules check for endings, some are lexicons looking for specific stems. The part of speech tags are assigned according to the following tagging strategies (the invocation order is mostly the one listed):

- whenever a word form is conclusively tagged and segmented, prefix a plus-sign to it marking that the remaining rule modules will not analyse this form; the final module removes all plus-signs;
- initially mark all potentially monosyllabic first syllables with a temporary diacritic (for facilitating ending identification; the final module removes all remaining diacritics);
- the top 200 word forms of the Finnish frequency dictionary are tagged and segmented as wholes;
- those 600 most common adverbs are segmented as wholes

that would otherwise be oversegmented by the ending procedures (note: lots of contemporary adverbs are petrified, etymologically discernible nominals);

- (the stems of) closed lexical classes are identified by lexical lists;
- segment all endings in a specified order dispersed over most of the 13 modules (this is the bulk of the morphology and largely equivalent to the work reported in Brodda & Karlsson 1981); those comparatively few unitary words that are potentially homonymous to endings are mostly listed and thereby exempted from segmentation;
- whenever possible, predict the part of speech labels on the basis of segmented inflexional or derivational endings;
- use (inconclusive) frequency-based stem-lexicons for adjectives and verbs to assign part of speech labels;
- by default, predict that the remaining untagged word forms are nouns.

Appendix 2 shows the intermediate phases in the process of tagging two sentences. Only six of the thirteen rule modules have been active. It is clearly seen how most of the "tagging load" resides with module 4 (the frequent word forms), module 6 (most verb endings; note: VI3 = 3rd infinitive, VPA2 = 2nd active participle, VPP2 = 2nd passive participle), module 9 (most nominal and some verb endings), module 12 (certain pronominal, numeral, and adjectival stems), and module 13 (by default, nouns).

The output contains two errors. PARI=A is, here, a numeral and not a noun. ESITELMINÄ is undersegmented, the correct analysis would be N:ESITELM=I=NÄ, an essive pl. left unsegmented on purpose due to surface homonymy with several frequent nouns with a base form ending in -ina, -inä. Such errors are, of course, due to the necessary but fallible heuristics inherent in any model not utilizing a total lexicon. Where conclusive decisions cannot be made, the most likely route is picked. In such instances, the heuristic choices have been made on the basis of large corpus studies.

So far, FINTAG has been applied to a text consisting of 66,000 word forms. It has an average success rate of some 85 % (in regard to word form tokens). A tagged form is taken to be correct only if no changes have to be made, i.e. the part of speech label is proper and disambiguated in context, all endings (if any) are properly segmented, and base forms are left unsegmented. Over- and undersegmentations are counted as errors, as are unresolved homonymies (e.g. N/A:SUOMALAINEN) and, of course, (eventual other) improperly placed boundaries and improper part of speech labels.

It deserves to be stressed that the success rate 85 % has been achieved on the level of word forms alone, without (a) active syntactic checking of the properties of neighbouring words or the whole clause, and (b) "memory" of previous decisions. This is a further proof of the high information load of overt morphological markers (also cf. Brodda & Karlsson 1981). Tentative experiments indicate that the success rate of FINTAG may fairly easily be raised to 93-95 % by invoking simple syntactic environment tests. In particular, this would provide effective means for disambiguating part of speech homonymies such as N/A:SUOMALAINEN, which, in the current version, constitute a large share of the "errors". Here we are approaching procedures akin to genuine parsing.

Note, for the sake of comparison, that Leech, Garside & Atwell (1983, 13) report a success rate of 96.7 % in tagging the LOB corpus. This figure is only minimally ahead of the morphologically dominated (revised) FINTAG.

The final version of FINTAG will be expanded to cover grammatical functions (subject, object, etc) and head-modifier relations, cf. the work done by Svartvik, Eeg-Olofsson, Forsheden, Oreström & Thavenius (1982) in syntactically tagging the Survey of Spoken English corpus. Even such information is, in Finnish, to a very large extent (over 90 %) inferrable from surface morph configurations. This is work in progress, based on LISP. Here, the aims and problems of tagging and parsing converge. Sophisticated "tagging" is nothing but applying theoretical "parsing" models.

## REFERENCES

- Bever, T.G. 1970. The cognitive basis for linguistic structures. In J.R. Hayes (ed.), *Cognition and the development of language*, John Wiley & Sons, N.Y., 279-352.
- Brodda, B. 1982. Problems with tagging - and a solution. *Nordic Journal of Linguistics* 5:2, 93-116.
- (forthcoming) The BETA system. *Data Linguistica*, Gothenburg.
- Brodda, B. & Karlsson, F. 1981. An experiment with automatic morphological analysis of Finnish. Department of General Linguistics, University of Helsinki, Publications No. 7.
- Karlsson, F. 1983. *Suomen yleiskielen äänne- ja muotorakenne*. WSOY, Porvoo.
- Kimball, J. 1973. Seven principles of surface structure parsing in natural language. *Cognition* 2, 15-47.
- Koskenniemi, K. 1983. Two-level morphology. A general computational model for word-form recognition and production. Department of General Linguistics, University of Helsinki, Publications No. 11.
- Leech, G., Garside, R. & Atwell, E. 1983. The automatic grammatical tagging of the LOB corpus. *ICAME News* No. 7, 13-33.
- Lieber, R. 1981. On the organization of the lexicon. IULC.
- Svartvik, J. & Eeg-Olofsson, M. & Forsheden, O. & Oreström, B. & Thavenius, C. 1982. *Survey of Spoken English*. Report on Research 1975-1981. CWK Gleerup, Lund.
- Winograd, T. 1983. *Language as a cognitive process*. Vol.1: Syntax. Addison Wesley, N.Y.

APPENDIX 1. Maximal set of derivatives for the verb *hakkaa* 'hew'

hakkaile	(a)	hakkauksellisin	hakkaistavuus
hakkailutta		<u>hakkauksellisimmuus</u>	hakkaistu
hakkautta		hakkaileminen	hakkaisematon
hakkaise		hakkailija	hakkaisemattomuus
hakkautu		hakkailijuus	hakkaisevainen
<u>hakkaantu</u>		hakkailijamainen	hakkaisevaisuus
hakkaaminen		hakkailijamaisuus	<u>hakkaus</u>
hakkaaja	(b)	hakkaileva	hakkautuminen
hakkaajuus		hakkaillevuus	hakkautuja
hakkaajatar		hakkailut	hakkautujuus
hakkaajattaruus		hakkailleisuus	hakkautujamainen
hakkaajamainen		hakkailtava	hakkautujamaisuus
hakkaajamaisuus		hakkailtavuus	hakkautuva
hakkaajatarmainen		hakkailtu	hakkautuvuus
hakkaajatarmaisuus		hakkaillematon	hakkautunut
hakkaajamaisempi		hakkailleemattomuus	hakkautuneisuus
hakkaajamaisemmuus		hakkaillevainen	hakkauduttava
hakkaajamaisin		hakkaillevaisuus	hakkauduttavuus
hakkaajamaisimmuus		<u>hakkailu</u>	hakkauduttu
hakkaava		hakkailuttaminen	hakkautumaton
hakkaavuus		hakkailuttaja	hakkautumattomuus
hakkaavampi		hakkailuttajuus	hakkautuvainen
hakkaavammuus		hakkailuttajamainen	<u>hakkautuvaisuus</u>
hakkaavin		hakkailuttajamaisuus	hakkaantuminen
hakkaavimmuus		hakkailuttava	hakkaantuja
hakannut		hakkailuttavuus	hakkaantujuus
hakanneisuus		hakkailuttanut	hakkaantujamainen
hakanneempi		hakkailuttaneisuus	hakkaantujamaisuus
hakanneemmuus		hakkailutettava	hakkaantuva
hakannein		hakkailutettavuus	hakkaantuvuus
hakanneimmuus		hakkailutettu	hakkaantunut
hakattava		hakkailuttamaton	hakkaantuneisuus
hakattavuus		hakkailuttamattomuus	hakkaannuttava
hakattavampi		hakkailuttavainen	hakkaannuttavuus
hakattavammuus		<u>hakkailuttavaisuus</u>	hakkaannuttu
hakattavin		hakkauttaminen	hakkaantumaton
hakattavimmuus		hakkauttaja	hakkaantumattomuus
hakattu		hakkauttajajuus	hakkaantuvainen
hakattuus		hakkauttajamainen	hakkaantuvaisuus
hakatumppi		hakkauttajamaisuus	
hakatummuus		hakkauttava	
hakatuin		hakkauttavuus	
hakatuimmuus		hakkauttanut	
hakkaamaton		hakkauttaneisuus	(a) = root → der. V
hakkaamattomuus		hakkautettava	(b) = root → der. N, A
hakkaamattomampi		hakkautettavuus	(c) = hakka/ile → der. N, A
hakkaamattomammuus		hakkautettu	(d) = hakka/il/utta → der. N, A
hakkaamattomin		hakkauttamaton	(e) = hakka/utta → der. N, A
hakkaamattomimmuus		hakkauttamattomuus	(f) = hakka/ise → der. N, A
hakkaavainen		hakkauttavainen	(g) = hakka/utu → der. N, A
hakkaavaisuus		<u>hakkauttavaisuus</u>	(h) = hakka/antu → der. N, A
hakkaavaisempi		hakkaiseminen	
hakkaavaisemmuus		hakkaisija	
hakkaavaisin		hakkaisijuus	(f)
hakkaavaisimmuus		hakkaisijamainen	
hakkaus		hakkaisijamaisuus	
hakkauksellinen		hakkaiseva	
hakkauksellisuus		hakkaisevuus	
hakkauksellisempi		hakkaissut	
hakkauksellisemmuus		hakkaisseisuus	
		hakkaistava	

APPENDIX 2. Intermediate phases of the tagging process. Numbers refer to the output of the respective active module. The effects of each module are underlined. Two incorrect outputs are starred. Cf. the text.

TÄ2MÄN  
 KO' KOELMAN  
 KI'RJOITUKSET  
 O' VAT  
 PA'RIA  
 KO'LMEA  
 LU'KUUN  
 O'TTAMATTA  
 SY'NTYNEET  
 VII'DEN  
 VII'ME  
 VUOLDEN  
 AI'KANA.  
 E'RÄÄT  
 NII2STÄ  
 +VF:ON  
 JU'LKISTETTU  
 LE'HDISTÖSSÄ,  
 E'RÄÄT  
 RA'DIOSSA,  
 E'RÄÄT  
 E'SITELMINÄ.

1

+PR:TÄMÄ=N  
 KO' KOELMAN  
 KI'RJOITUKSET  
 +VF:O=VAT  
 PA'RIA  
 KO'LMEA  
 LU'KUUN  
 O'TTAMATTA  
 SY'NTYNEET  
 VII'DEN  
 +A:VIIME  
 +N:VUODE=N  
 +N:AIKA=NA.  
 E'RÄÄT  
 +PR:NI=I=STÄ  
 +VF:ON  
 JU'LKISTETTU  
 LE'HDISTÖSSÄ,  
 E'RÄÄT  
 RA'DIOSSA,  
 E'RÄÄT  
 E'SITELMINÄ.

4

+PR:TÄMÄ=N  
 KO' KOELMAN  
 KI'RJOITUKSET  
 +VF:O=VAT  
 PA'RIA  
 KO'LMEA  
 LU'KUUN  
 +VI3:OTTA=MA=TTA  
 +VPA2:SYNTY=NEE=T  
 VII'DEN  
 +A:VIIME  
 +N:VUODE=N  
 +N:AIKA=NA.  
 E'RÄÄT  
 +PR:NI=I=STÄ  
 +VF:ON  
 +VPP2:JULKISTE=TTU  
 LE'HDISTÖSSÄ,  
 E'RÄÄT  
 RA'DIOSSA,  
 E'RÄÄT  
 E'SITELMINÄ.

6

+PR:TÄMÄ=N  
 KOKOELMA=N  
 KIRJOITUKSE=T  
 +VF:O=VAT  
 PARI=A  
 KOLMEA  
 LUKU&UN  
 +VI3:OTTA=MA=TTA  
 +VPA2:SYNTY=NEE=T  
 VIIDE=N  
 +A:VIIME  
 +N:VUODE=N  
 +N:AIKA=NA.  
 ERÄÄ=T  
 +PR:NI=I=STÄ  
 +VF:ON  
 +VPP2:JULKISTE=TTU  
 LEHDISTÖ=SSÄ,  
 ERÄÄ=T  
 RADIO=SSA,  
 ERÄÄ=T  
 ESITELMINÄ.

9

+PR:TÄMÄ=N  
 KOKOELMA=N  
 KIRJOITUKSE=T  
 +VF:O=VAT  
 PARI=A  
 +NUM:KOLME=A  
 LUKU&UN  
 +VI3:OTTA=MA=TTA  
 +VPA2:SYNTY=NEE=T  
 +NUM:VIIDE=N  
 +A:VIIME  
 +N:VUODE=N  
 +N:AIKA=NA.  
 +PR:ERÄÄ=T  
 +PR:NI=I=STÄ  
 +VF:ON  
 +VPP2:JULKISTE=TTU  
 LEHDISTÖ=SSÄ,  
 +PR:ERÄÄ=T  
 RADIO=SSA,  
 +PR:ERÄÄ=T  
 ESITELMINÄ.

12

Final output

PR:TÄMÄ=N  
 N:KOKOELMA=N  
 N:KIRJOITUKSE=T  
 VF:O=VAT  
 ★N:PARI=A  
 NUM:KOLME=A  
 N:LUKU&UN  
 VI3:OTTA=MA=TTA  
 VPA2:SYNTY=NEE=T  
 NUM:VIIDE=N  
 A:VIIME  
 N:VUODE=N  
 N:AIKA=NA.  
 PR:ERÄÄ=T  
 PR:NI=I=STÄ  
 VF:ON  
 VPP2:JULKISTE=TTU  
 N:LEHDISTÖ=SSÄ,  
 PR:ERÄÄ=T  
 N:RADIO=SSA,  
 PR:ERÄÄ=T  
 ★N:ESITELMINÄ.

13